Corpus Development for Affective Video Indexing

Mohammad Soleymani, *Member, IEEE,* Martha Larson, *Member, IEEE,* Thierry Pun, *Member, IEEE,*
Alan Hanjalic, *Senior Member, IEEE,*

Abstract—Affective video indexing is the area of research that develops techniques to automatically generate descriptions that encode the emotional reactions which videos evoke in viewers. This paper provides a set of corpus development specifications based on state-of-the-art practice intended to support researchers in this field. Affective descriptions can be used for video search and browsing systems offering users affective perspectives. The paper is motivated by the observation that affective video indexing has yet to fully profit from the standard corpora (data sets) that have benefited conventional forms of video indexing. Affective video indexing faces unique challenges, since viewer-reported affective reactions are difficult to collect, and collection efforts must be carefully designed in order to both cover the full scope of affective response and also capture its stability. We first present background information on affect and multimedia and related work on affective multimedia indexing, including existing corpora. Three dimensions emerge as critical for affective video corpora, and form the basis for our proposed specifications: the context of viewer response, personal variation among viewers, and the effectiveness and efficiency of corpus creation. Finally, we present examples of three recent corpora and discuss how these corpora make progressive steps towards fulfilling the specifications.

Index Terms—Emotional characterization, benchmarks, multimedia, content analysis, videos

1 Introduction

Video indexing is the process of analyzing video content in order to extract a representation that is specific enough to characterize the uniqueness of the content and, at the same time, is abstract enough to capture useful similarities with other video content. Research and development in the area of video indexing falls under the larger domain of multimedia content analysis, which includes the theories, algorithms and systems that extract or infer descriptors which encode characteristics of multimedia content. These descriptors take a variety of forms, ranging from machine interpretable indexing features, to metadata labels in the form of textual words or phrases that can also be interpreted directly by humans (e.g., [1], [2], [3]). The common function of such descriptors is to represent video content in a way that enables the implementation of systems that give users better access to multimedia content. In particular, here, we are interested in video indexing techniques that will be used for video search engines and other systems that support browsing video collections or otherwise representing to users the contents of a video stream.

Conventionally, video indexing has focused on describing videos in terms of the content that humans identify as being explicitly depicted in their visual channel. Much attention has been devoted to developing algorithms that detect visual concepts in video that are related to events, objects, people, scenes, and locations [4]. Such concepts can be considered the 'literal' content of a video. The meaning or the value of a particular video for a viewer clearly goes far beyond its literal content however. Videos can also be characterized in terms of how they influence viewers' emotions, i.e., their affective impact on viewers. Affective impact refers to the intensity and type of emotion that is evoked in a viewer while watching a video. The potential of affective indexing to contribute to the automatic creation of descriptions that are useful for video search engines is widely acknowledged. However, much research in the area of video indexing remains focused on literal descriptions of video and affective video indexing has yet to reach its full potential.

An important factor contributing to the success of visual concept detection and other literal approaches to indexing video is the existence of standardized corpora (data sets). These corpora are made available to the research community, often within the framework of a benchmarking initiative, and can be used by researchers to evaluate the algorithms that they develop. For example, detection of visual concepts in video have been a primary focus for the largest multimedia benchmarking efforts, most notably TRECVid [5], [6]. Similar large, high-quality data sets, used at the community level in benchmarking initiatives, have yet to be developed for affective video indexing.

- *Mohammad Soleymani is with the Intelligent Behaviour Understanding Group (iBUG), Imperial College London, 180 Queen's Gate, London SW7 2AZ, United Kingdom. (Email: m.soleymani@imperial.ac.uk).*
- *Martha Larson and Alan Hanjalic are with the Multimedia Information Retrieval Lab, Delft University of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands. (Email: {m.a.larson,a.hanjalic}@tudelft.nl).*
- *Thierry Pun is with the Computer Vision and Multimedia Laboratory, University of Geneva, Battelle Campus, Building A, Rte. de Drize 7, Carouge(GE) CH - 1227, Switzerland. (Email: thierry.pun@unige.ch).*

This paper takes the position that corpora have a key role to play in supporting the research work that is necessary in order to allow affective video indexing to reach its full potential. The main contribution of this paper is a set of 'affective video indexing corpus development specifications' that arise from a discussion of the state of the art and an analysis of the limitations of existing data sets. The specifications are organized along three dimensions that are identified as critical for the process of corpus development for affective video indexing: the context of viewer response, personal variation among viewers, and the effectiveness and efficiency of the process of collecting viewer-reported affective reactions.

The paper is organized as follows. In the remainder of this section, we set the scene, motivating affective video indexing research and discussing how corpus development contributes to its advancement. Section 2 provides background material on affect in multimedia and discusses existing techniques. Then, Section 3 covers previous work on affective video indexing, and the corpora that have been used in this work. Building on the information in Section 2 and 3, we formulate a set of corpus development specifications for affective video indexing, which we present in Section 4. Next, in Section 5, we introduce a series of three corpora that we have developed using three different settings for the collection of viewer affective response: the laboratory, a Web-based online platform and a crowdsourcing platform. These corpora illustrate progressively more advanced applications of our proposed corpus development specifications. We finish in Section 6 with conclusions and an outlook on the future of corpus development for affective video indexing.

1.1 The rise of affective video indexing

The affective video indexing paradigm assumes that users' needs for multimedia content involve a strong affective component and that a multimedia information system, e.g., a video search engine, must be able to take feelings, emotion and mood into account. Recently, the awareness of the importance of affect in people's information seeking behavior has been growing, as witnessed by work in the area of conventional text information retrieval, such as [7], [8]. In parallel, awareness of the potential of affective indexing for multimedia information retrieval has also increased.

Affective video retrieval was first discussed in the mid-1990's by Rosalind Picard as an application of affective computing [9]. Affective video indexing is well summarized by her statement, "Although affective annotations, like content annotations, will not be universal, they will still help reduce time searching for the 'right scene'." [9] (p. 11). When it was first introduced, the proposal that affect could provide an effective means to organize video was not immediately widely accepted. However, a decade later, the idea had matured in form and established its status as a new paradigm within multimedia information retrieval community [10].

The importance of affect is now widely accepted by researchers, as reflected by [11], a survey of multimedia information retrieval, which states that "On a fundamental level, the notion of user satisfaction is inherently emotional." (p. 3). The current paper is motivated by our conviction that the availability of large, high-quality corpora for the evaluation of affective video indexing will support the multimedia research community in turning its awareness of the importance of affective video indexing into tangible and significant advancement of the state of the art.

1.2 The challenge of affective video indexing

The central challenge faced by affective video indexing lies in the difference between descriptions that refer to the affective impact of videos and descriptions that refer to the literal content of the video. In the case of descriptions of literal content, viewers can quickly and consistently assess or confirm whether a description is relevant for a given video, e.g., whether or not a given visual concept is depicted in the video. In making this judgment, they rely on cognitive processing combined with general world knowledge. The judgment is considered to be objective because it can be easily reproduced by consulting a group of viewers, largely independently of the viewers' backgrounds.

Assessing the affective impact of a video, i.e., collecting affective descriptions of video, is less clear cut. Information on affect can be gathered by asking viewers to report their emotional response upon watching the video. Affective response is considered to be subjective, since only the subject experiencing the response (i.e., the viewer) is in a position of authority to assess or confirm a particular response. It is tempting to conclude that subjectivity (i.e., the fact that no observer other than the viewer has access to direct knowledge of the viewer's affective response) makes the problem of predicting affective response to a video hopelessly ill defined. Indeed, the affective response evoked in a viewer while watching the video is personal in that it can, and does, differ from person to person. It is also contextual, since it varies when the context in which the video is watched or the underlying mood or physical state of the viewer changes.

However, although it is not clear cut, affective response is far from arbitrary. In many cases, affective impact will be quite consistent and there will be a high level of agreement in affective response across viewers. The challenge of affective indexing for video involves how to identify those aspects of video that trigger emotional reactions across viewers that are stable enough that they can be robustly predicted.

The stability of affective impact is most clearly illustrated in the case of film. Filmmakers are highly skilled in evoking specific emotions in their audiences. The high-level of inter-subjective agreement concerning the connotative aspects of film has been studied and used as the basis for an automatic indexing system by [12].

Connotation is that dimension of interpretation that goes beyond literal meaning, and, as such, encompasses a large affective component. Today's video search engines index large quantities of video on the Web. For this video, it is not possible to make a priori assumptions about the extent to which the techniques and conventions used in formal film to trigger affective response in viewers apply. However, it is possible to anticipate that there will be a component of viewer response that is grounded in modalities of emotional reaction shared in the audience or arising from common interpretation conventions.

This paper takes the standpoint that by isolating and emphasizing aspects of video for which human judges display a relatively high level of agreement, corpora for the evaluation of affective indexing can be created that can make a contribution to advancing the state of the art comparable to the contribution made by benchmarks that focus on literal descriptions of video content.

1.3 The contribution of corpus development

Corpus development contributes to advancing the state of the art of multimedia technology by making possible standardized evaluation. Only when a standard data set and ground truth are used, is it possible to directly and fairly compare alternative algorithms. Comparison and reproducibility help to drive forward the state of the art: when researchers know how their algorithms perform with respect to the state of the art they can better direct their efforts to surpass it and more quickly abandon less promising lines of investigation. Benchmarks and standard tasks/data sets help to eliminate redundancy by enabling direct comparison between algorithms across research sites, increasing the efficiency of the research community by allowing resources to be shared between sites and providing a framework in which researchers can interact in a mixture of collaboration and competition that is stimulating and productive. The impact of the TRECVid evaluation for video has been large and is well documented [13]. However, as mentioned above, TRECVid focuses on literal approaches to video indexing, i.e., content explicitly depicted in the visual channel. Corpus development is key to allowing affective video indexing to achieve similar impact.

Corpus development does not strive to promote one particular variety of affective video indexing, but rather if numerous, well-designed multimedia corpora were available, they would contribute in many different ways. Here, we provide some examples of the range of applications in which affective video indexing, retrieval and browsing has been used. In [14], highlights were extracted from baseball programs and in [2] an adaptive approach to sports video highlight detection was proposed and studied in detail for the case of soccer. Retrieval of movie clips using multimedia content features and user-assigned keywords was investigated by [3]and [15]. Laughter events have been successfully deployed in videos for navigation [16]. The examples illustrate the spread of application areas that stand to benefit if large, high-quality corpora can be developed in made available to the research community.

Currently, data sets used to evaluate individual theories and algorithms in affective content analysis are typically limited in size and scope. The limitations are imposed because of the relative difficulty of collecting affective responses from many viewers. These limitations also reduce variability in the elicited affective responses of test users, which facilitates manual annotation and results interpretation, but may ultimately be too narrow for the resulting algorithms to be used in practical situations.

Recently, however, technological developments have provided means for developing a new generation of corpora. Online systems make it possible to ask large numbers of viewers to watch videos and provide information on their affective response. Additionally, the rise of crowdsourcing and large crowdsourcing platforms such as Amazon Mechanical Turk (www.mturk.com) make it possible to more easily recruit large numbers of annotators with a representative spread of backgrounds. Corpora in existence today do not, as yet, fully exploit these resources. The corpus development specifications set out in this paper aim to encourage the effective use of the new opportunities offered by the Web and by crowdsourcing platforms.

2 Background and Existing Techniques

This section provides an explicit specification of the key concepts of affect and multimedia that are used in this paper and covers the relevant related work. First, we provide a clear definition of affective viewer response to multimedia as it is applied for affective video indexing and discuss this definition with respect to the larger field of research on human emotion. Then, emotional representations and existing tools that have been developed to collect annotations in the form of these representations are introduced.

2.1 Emotion in response to multimedia

Emotions are complex phenomena with affective, cognitive, conative and physiological components [17]. The affective component is the subjective experience conventionally connected with feelings. The cognitive component is the perception and evaluation of the emotional situation. The conative component is the expression of affect, including facial expressions, body gesture, and any other action that has a preparatory function for an emotional situation. The physiological component regulates physiological responses in reaction to the emotional situation, for example, increasing perspiration during a fearful experience. When studying emotion evoked in viewers in response to multimedia, it is important to take the complexity of emotion into account rather than expecting emotion to manifest itself along a single dimension only.

In understanding emotional response, the terms "mood" and "emotion" should be differentiated. We mention this point explicitly, since these terms are sometimes used interchangeably in the literature despite the clear formal distinction between their definitions. Mood is a diffused affective state that is long, slow moving and not tied to a specific object or elicitor whereas emotions can occur in short moments with higher intensities [18]. Scherer defines this by intrinsic and extrinsic appraisal [18]. Intrinsic appraisal is independent from the current goals and values of the viewer while extrinsic or transactional appraisal leads to feeling emotions in response to the stimuli. For example, the intrinsic emotion of an image depicting someone who is smiling is happiness. If the person smiling is a figure disliked by the viewer, extrinsic appraisal leads to unpleasant emotions. In this paper, we are concerned with extrinsic emotion, with video as the elicitor of the emotional response.

One of the most well-known and widely-accepted theories that explains the development of emotional experience is appraisal theory. According to this theory, cognitive judgment or appraisal of a situation is a key factor in the emergence of emotions [19], [20], [21]. According to Orthony, Clore and Collins (OCC) [20], emotions are experienced following a scenario comprising a series of phases. First, there is a perception of an event, object or an action. Then, there is an evaluation of events, objects or action according to personal wishes or norms. Finally, perception and evaluation result in a specific emotion of emotions arising. Considering this scenario for an emotional experience in response to multimedia content, emotions arise first through sympathy with the presented emotions in the content [17]. During the appraisal process for an emotional experience in response to multimedia content, viewers examine events, situations and objects with respect to novelty, pleasantness, goal, attainability, copability, and compatibility with their norms. Then, the viewers' perceptions induce specific emotions, which changes their physiological responses, motor actions, and feelings.

Emotional processes can be divided into different categories. Here, we mention three processes that apply not only in the general case of emotional response, but also in the specific case of viewer affective response to video: *emotion induction*, *emotional contagion* and *empathic sympathy* [17]. An example of *emotion induction* is when in a TV show a politician's comment makes the viewers angry while the politician himself is not angry. The angry response from the viewers is due to their perception of the situation according to their goals and values. *Emotional contagion* occurs when the viewer only perceives the expressed emotion from a video. For example, the induced joy as a result of sitcom laughter can be categorized in this category. In the empathic category, the situation or event does not affect the viewer directly, but rather the viewer reproduces the appraisal steps of the characters who are depicted in the video. The empathic reaction is called *symmetric co-emotion* in case the viewer has positive feelings about the character and *asymmetric co-emotion* in case the viewer has negative feeling about the character [22].

Empathy is a complex phenomenon that has both cognitive and affective components. Affective empathy is the primitive response involved in sympathizing with other individuals. On the other hand, cognitive empathy is the intellectual understanding of other people and the rational reconstruction of their feelings [23], [17]. Zillman developed an affective disposition theory for narrative plot [24], [22]. According to this theory, empathic emotions originate with the observation of the actors by viewers. First, a character's actions are morally judged by the viewer and the judgment results in a positive or negative perception of the character. Then, depending on whether the viewer approves or disapproves of the character, the viewer sympathizes either empathically or counter-empathetically. The intensity of the perceived emotion in response to a video depends on how much viewers identify themselves with the protagonists and to what extent they suspend their own identities while watching the video [24].

In general, it is a daunting challenge to go from appraisal theory and the sources of affective empathy to a technique that analyzes the video signal to predict viewer effective response. However, some characteristics of video are quite indicative of affective response. Much video will capture not only action, but the audience of that action. This audience might be the spectators of a sports event or certain characters in a film, who watch and react to events. The reaction of these in-video observers (e.g., laughter or cheering) can provide important clues to how viewers will react to the video. Further, the literature has identified the most important emotion inducing components of movies as being music and narrative structures [17]. Music is clearly instantiated at the signal level, and structure can also to a certain extent be extracted (e.g., the quick shot changes of a chase scene that will correspond to abrupt changes in the visual constitution of a scene). In pursuit of such regularities, researchers have undertaken to develop techniques for affective video content analysis, which we turn to discuss in Section 3.

2.2 Emotional representations

Different emotional representations have been developed by past research, including, discrete, and continuous models. Discrete emotions theories are inspired by Darwin and support the idea of the existence of the certain number of basic and universal emotions [18], [25]. Darwin suggested that emotions exist because they are important for survival. Different psychologists proposed different lists of basic emotions. The so called basic emotions are mostly utilitarian emotion and their number is usually from 2 to 14. Scherer also proposed a list of emotional keywords to code discrete and free choice emotional reports [18].

Wundt [26] was the first to propose a dimensional representation for emotions. Dimensional theories of emotion suggest that emotions can be represented as points in a continuous space and discrete emotions are folk-psychological concepts [27].

Discrete emotions also present a challenge for representation. One particularly important aspect of this challenge is that keywords are not cross-lingual. In other words, emotions do not have exact translations in different languages, e.g., there is no word in Polish that corresponds exactly in meaning to the English word, "disgust" [28].

Psychologists often represent emotions in an n-dimensional space (generally 2- or 3-dimensional). The most well-known example of such a space arises is the 3D valence-arousal-dominance or Pleasure-Arousal-Dominance (PAD) space [29]. This space arises from cognitive theory and is widely used for studying affect and multimedia—we ourselves make use of it for corpus developed, as discussed later in the paper. The valence scale ranges from unpleasant to pleasant. The arousal scale ranges from passive to active or excited. The dominance scale ranges from submissive (or "without control") to dominant (or "in control, empowered"). Fontaine et al. [30] proposed adding predictability dimension to PAD dimensions. Predictability level describes to what extent the sequence of events is predictable or surprising for a person.

2.3 Emotional self-reporting methods

Understanding the "true", underlying emotion that was felt by a participant during an experiment has been always a challenge for psychologists. Multiple emotional self-reporting methods have been created and used so far [31], [32], [18], [33], [34]. Emotional self-reporting can be done either in free-response or forced-choice formats. In the free-response format, the experiment participants are free to express their emotions by words. In the forced-choice, participants are asked to answer specific questions and indicate their emotion. Forced-choice self-reports in affective experiments use either discrete or dimensional approaches. Based on discrete emotions, self-reporting tools have been developed that can be used to ask participants to report their emotions with emotional words on nominal and ordinal scales. Dimensional approaches of emotional self-reporting are based on bipolar dimensions of emotions. Emotions can be reported along each dimension using ordinal or continuous scales [33]. Here, we discuss in more detail some popular self-reporting methods which have been used for psychological and human computer interaction research.

Russell [35] introduced the "circumplex model" of affect for emotion representation. In his model, eight emotions; namely, "arousal", "excitement", "pleasure", "contentment", "sleepiness", "depression", "misery" and "distress" are positioned on a circle surrounding a two dimensional activation, pleasure-displeasure space. Starting form these eight categories, 28 emotional keywords were positioned on this circumplex, based on the results of a user study. The advantage of this circumplex over either discrete or dimensional models is that all the emotions can be mapped on the circumplex using only the angle. In this way, all emotions are presented on a circular and one dimensional model.

The Self Assessment Manikin (SAM) is one of the most famous emotional self-reporting tools. It consists of manikins expressing emotions. The emotions vary along three different dimensions; namely, arousal, valence, and dominance [33]. The SAM Manikins are shown in Fig. 1. Experiment participants can choose the manikin that best portrays their emotion. This method does not require the verbalization of emotions and the manikins are understandable without further explanation. For these reasons, the SAM tool is language independent. The second advantage of the SAM tool is that it can be directly used in measuring the multiple dimensions of emotions. A limitation of SAM is that subjects are unable to express co-occurring emotions with this tool.

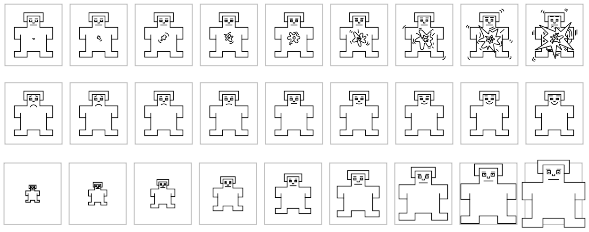


Fig. 1. Self Assessment Manikins. From top to bottom the manikins express different levels of arousal, valence, and dominance.

The "Positive and Negative Schedule" (PANAS) [36] permits self-reporting 10 positive and 10 negative affects on a five-point scale. An expanded version of PANAS, the "Positive and Negative Schedule—Expanded Form" (PANAS-X), was developed later. PANAS-X provides the possibility of reporting 11 discrete emotion groups on a five-point scale [37]. PANAS is made to report affective states and can be used to report both moods and emotions. PANAS-X includes 60 emotional words and takes on average 10 minutes for an experimental participant to complete [37]. The time needed to answer the PANAS questionnaire makes it too difficult to use in the experiments with limited time and multiple stimuli.

Scherer [18] positioned 20 emotions around a circle to combine both dimensional and discrete emotional approaches, and in this way created the Geneva emotion wheel. For each emotion around the wheel, five circles whose size increases from the center outwards are displayed. The size of the circles is an indicator of the intensity of felt emotion (see Fig. 2). In an experiment, participants can pick, from the list of 20 emotions, up to two emotions that were the closest to their experience and report the intensities of the emotions with the size of the marked circles. In case, no emotion is felt, a user can mark the upper half circle in the hub of the wheel. If a different emotion is felt by a user, it can be indicated in the lower half circle. The emotions are sorted on the circle such that, high-control emotions are on the top and low-control emotions are at the bottom and the horizontal axis, which is not explicitly visible on the wheel, represents valence or pleasantness.

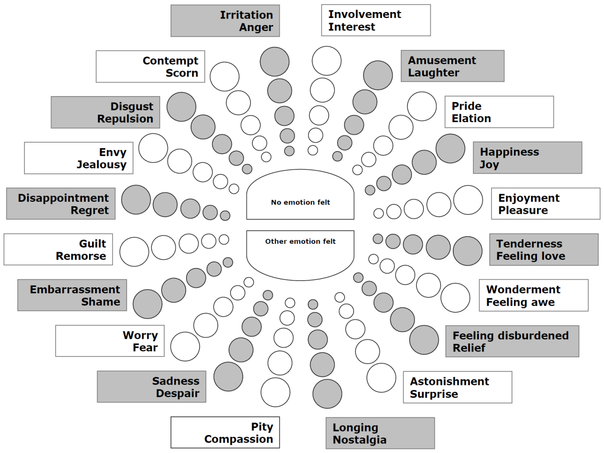
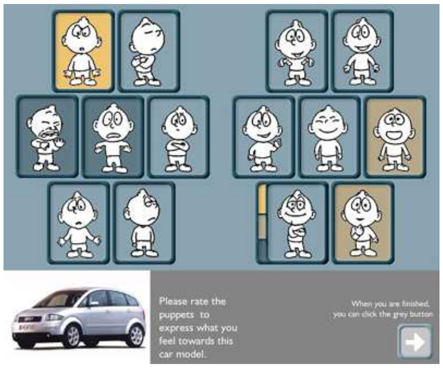


Fig. 2. Subjects can indicate their emotion on the Geneva emotion wheel by selecting the corresponding circle.

PrEmo is an alternative non-verbal emotion reporting tool to report emotions in response to product design. Desmet proposed PrEmo to overcome the problem of reporting co-occurring emotions by making use of animated characters expressing emotions [31]. PrEmo consists of 14 animated characters expressing different emotions and it is, for this reason, language independent. Users can assign a score, at three levels, to one or more characters that they identify as relevant to their emotional response (see Fig. 3).


Please rate the puppets to express what you feel towards this car model.

When you are finished, you can click the grey button

Fig. 3. Users can identify emotion they are feeling with 14 animated characters are expressing emotions.

2.4 Video affective annotation tools

Among existing self-reporting tools, few have been designed specifically for the affective annotation of video. Villon developed an annotation tool with which a user can drag and drop videos onto the valence-arousal plane [38]. This tool presents the possibility of comparing the ratings given to different videos and enables an experiment participant to rate a video relative to the ratings given other videos. The tool enables users to take their previous reports into account while annotating a new video.

Feeltrace was developed to annotate the intrinsic emotion in videos [39]. This tool is originally designed to annotate the emotions expressed by people who are depicted in videos (e.g., in talk shows), including acted facial expressions or gestures [40]. Although this tool provides the possibility of continuous annotation, it is not an appropriate tool for emotional self-reporting, because it is difficult for tool users to both concentrate on the video and reporting changes in their emotions.

An online video affective annotation tool has been developed by Soleymani et al. [41]. With their annotation tool, a experiment participant can self-report emotions after watching a given video clip by means of SAM manikins and emotional keywords from a selected list in a drop down menu. This tool is used in the development of our web-based corpus presented in Section 5.2.

3 Affective Video Indexing

In this section, we provide an overview of affective indexing including methods and discuss the corpora that have been developed in previous work. In particular, we discuss the shortcomings of the existing corpora, which are used as a basis to develop a set of specifications for the design and development of future corpora.

3.1 Affective video analysis for Indexing

Affective video content analysis involves estimating the affective response elicited in viewers by the content. Motivated by work in the area of film, researchers have extracted content features, such as audio energy and color histograms from the video signal and used machine learning techniques to infer which emotion would be felt by an average viewer. They have considered different goals and applications for their algorithms, from video summarization to personalized content delivery.

A summary of key examples from the existing literature in the area of content analysis for the emotional understanding of videos is given in Table 1. The table contains information about the data set that was used to evaluate the proposed algorithms, including the information about the type of representation used (discrete vs. continuous), the affective categories used, the number of human annotators used to create the corpus, the modalities contained in the video data set and finally, the results of evaluation, if they are given by the paper.

TABLE 1
Key examples of previous work on multimedia content analysis for affective video indexing and existing corpora.

Study	Emotion repr.	Categories or dimensions	Nr of Annotators	Modalities	Evaluation results
Kang [1]	disc.	fear/anger, joy, sadness and neutral	10	V	classification rate, fear: 81.3%, sadness: 76.5%, joy: 78.4%
Hanjalic & L.-Q. Xu [43]	cont.	valence and arousal	-	AV	no evaluation
Wang & Cheong [44]	disc.	fear, anger, surprise, sadness, joy, disgust and neutral	3	AV	74.7%
Arifin & Cheung [45]	cont.	pleasure, arousal, and dominance	14	AV	-
Xu et al. [46]	disc.	fear, anger, happiness, sadness and neutral	unknown	AV	80.7%
Soleymani et al. [47]	disc. & cont.	continuous arousal for shots, positive/negative excited & calm on scene level	1	AV & text	63.9%
Irie et al. [48]	disc.	acceptance, anger, anticipation, disgust, joy, fear, sadness, surprise and netural	16	AV	subject agreement rate 0.56
Joho et al. [49]	cont.	valence and arousal	10	AV	-
Teixeira et al. [50]	disc.	fear, anger, surprise, sadness, joy, disgust	6	AV	71.4%
Demarty et al. [42]	violence	violent & non-violent	7	AV	-

Existing approaches for content analysis generally make use of low-level content features extracted from both video and audio. Low-level audio features often include short time energy, zero crossing rate, Mel Frequency Cepstral Coefficients (MFCC), and pitch. Low level visual features include, color variance, motion component, shot change rate, key lighting, brightness, and color energy [46], [1], [43], [45]. Irie et al. [48] proposed using a bag of audio-visual words strategy to transform the feature space before classification.

Different machine learning models have been used to classify videos on different levels, e.g., shots, scenes, into different emotional classes, with the goal of emotional tagging. Kang [1] used a Hidden Markov Models (HMM) classifier to detect emotional events from low level features. Hanjalic and L.-Q. Xu [43] applied a regression model to predict arousal and valence on continuous temporal dimension. M. Xu et al. [46] proposed using a hierarchical approach that first clusters the samples in the arousal dimension and then classified them using HMMs into valence classes. Soleymani et al. [47] used movie genres and the temporal dimension to predict emotions of scene level using a Bayesian framework. Irie et al. [48] used a latent topic model by defining affective audio-visual words in the content of movies to detect emotions in movie scenes. This model takes into account temporal information, i.e., the effect of the emotion from the preceding scene, to improve affect classification. The probability of emotional changes between consecutive scenes was also used in [47] to improve emotional classification of movie scenes using content features.

Going beyond multimedia content analysis, emotional responses of the viewers have been also used to detect affective tags for videos. Joho et al. [49], [51] developed a video summarization tool using facial expressions. Kierkels et al. [52] proposed a method for personalized affective tagging of multimedia using peripheral physiological signals. Valence and arousal levels of participants' emotion when watching videos were computed from physiological responses using linear regression [53].

We next survey existing corpora and point out their strengths as well as their shortcomings that necessitate the development of new corpora.

3.2 Existing corpora

In this section, we provide a summary of corpora that have previously been developed and used for evaluating affective video indexing. In general, affective video corpora are developed with specific goals. Three common goals are: first, emotion elicitation or mood regulation in psychological experiments; second, emotional characterization of videos using content for video indexing or highlighting and third, recognition of the intrinsic emotions in the videos, e.g., detecting the emotions which were expressed by people in the videos. Although all three involve emotion, it is critical to avoid mixing these three different research tracks and the goals behind them. For example, movie excerpts that are most likely to elicit strong emotions are chosen for emotion elicitation in the case of the first goal. In contrast, for the second goal, an exclusive focus on strongly emotional excerpts is not appropriate for emotional characterization. Emotional characterization should be able to deal with the full spectrum of emotions in videos, from neutral videos to mixed and strong emotions.

Rottenberg et al. [54] created an emotional video dataset for psychological emotion elicitation studies. The excerpts, which were about 1–10 minutes long, were either extracted from famous commercial movies or from non-commercial videos that were used in emotional research, e.g., an amputation surgery video. First, they formed a set of excerpts with different targeted emotions; namely, amusement, anger, disgust, fear, neutral, sadness and surprise. They evaluated the excerpts based on "intensity" and "discreteness". The "intensity" of an excerpt means whether a video received high mean report on the target emotion in comparison to other videos. The "discreteness" refers to what extent the target emotion was felt more intensely in comparison to all non-targeted emotions using the ratings a video received on the target emotion in comparison to the other emotions. They ultimately formed a dataset consisting of 13 videos, from under a minute to up to eight minutes long, for emotion elicitation studies.

In a more recent study, Schaefer et al. [36] created a larger dataset from movie excerpts to induce emotions. In their study, they went beyond discrete basic emotions and developed a corpus including 15 mixed feelings in addition to six discrete emotions; namely, anger, disgust, sadness, fear, amusement, tenderness. 364 participants annotated their database using three questionnaires.

Almost all research work published in the field of multimedia content analysis and emotions has used its own individually developed corpus. Table 1 provides and overview of key examples of such work, specifying details of the affect categories used and the modalities of the video (i.e., audio and/or video) that were used to carry out the analysis. The table allows comparison of the corpora that were used to evaluate these techniques and the number of results. In the following, we provide additional details on this work, by discussing specific examples in greater depth.

Wang and Cheong [44] created and annotated a dataset consisting of 36 full length Hollywood movies which have 2040 scenes. Three annotators watched the movies and reported their emotions specifying Ekman basic emotion labels [55] for every scene. Only 14% of the scenes received double labels and the rest only received single emotional labels from their three annotators.

Hanjalic and Xu [43] used excerpts from the movies "Saving Private Ryan" and "Jurassic Park 3" and two soccer matches in their study without annotations. Irie et al. [48] used 206 selected emotional scenes out of 24 movies. A total of 16 students annotated these scenes using eight Plutchik basic emotions: "joy", "acceptance", "fear", "surprise", "sadness", "disgust", "anger", and "anticipation" [56]. The annotators first watched the videos and then reported how much they felt each of these emotions on seven points scale. The emotional labels were assigned to the selected scenes only if more than 75% of annotators agreed on them, otherwise the neutral label was assigned to the movie scene. M. Xu et al. [46] used selected scenes from eight movies containing 6201 shots which are in total 720 minutes long. The videos were manually labeled by five emotions: fear, anger, happiness, sadness and neutral spanning the arousal dimension in three levels and valence in two levels.

Soleymani et al. [47] used 21 full length commercially produced movies. One annotator annotated the movies continuously using an annotation tool which was recording the coordinates of mouse on valence and arousal plane on every click. The annotator reported his emotion at every moment he felt a different emotion while watching the movies.

Teixeira et al. [50] used selected excerpts from 24 movies. They first segmented the movies into short clips (M=112s), and showed them to 16 participants. Participants rated the movies using SAM Manikins [33] on a seven-point scale; 346 clips, 10h 26 min in total, were chosen to span arousal, valence, dominance space.

Demarty et al. [42] created a benchmark consisting of 18 Hollywood movies for violence detection. Although the movies are not annotated directly by emotional terms, depiction of violence elicits negative emotions. The dataset is annotated by seven annotators on shot level.

3.3 Open Issues with the existing corpora

We end this section with a summary of the limitations of currently existing corpora in the form of a catalogue of open issues. First, we have pointed out that the same viewer can experience different emotions in response to the same stimulus depending on the context. The importance of the influence of context for viewer affective response to video is relatively uncontroversial. For example, it is not strange or surprising when someone remarks, "I am not in the mood to watch that movie today." Our survey has revealed that the assumptions and methodology adopted by existing work is inconsistent with the importance of context for affective reactions. Researchers often emphasize controlling the context and conditions in which annotations are collected from users, or disregard the issue of context entirely when designing experiments and developing data sets. Ignoring or suppressing context introduces risk into affective video indexing research: a system that does not take into account the high degree of variability that characterizes naturally occurring contexts in which video is consumed may not be able to respond appropriately to user needs in real-world situations. As we will discuss further in Section 5, there are a wide variety of contextual dimensions with a significant effect on the emotions that viewers feel in response to a video, including time of the day, temperature, mood and social context.

Second, beyond contextual factors, most of the existing research on affective video characterization has assumed reactions to be homogeneous across viewers, e.g. [44]. In some cases, the assumption of a single, obvious affective reaction from viewers is so strong, that affective video analysis is carried out, without collecting any user annotations at all [43]. In most cases, however, assuming that everyone will react in the same way when watching a particular video is a strongly limiting assumption, that contradicts our intuition that the subjective nature of affect includes a strongly individual dimension. Corpora that allow both the personal and general dimensions of affective reactions to be explored, have greater potential in helping to advance algorithm development in a direction that will best cover the needs of the full spectrum of possible users.

Finally, in order to model affective responses that vary over context and across videos, affective video corpora are needed that include a large number of responses collected from a very large and representative population. However, the number of viewers and their feedback are often limited by our experimental setting and resources. In order to carry out research within the practical constraints of the real world, methods for creating affective video indexing corpora must be both effective—resulting in useful, high-quality corpora—and also efficient with respect to both the time spent and the expense incurred in the development process.

These open issues constitute three dimensions that inform our proposal of specifications for corpus development for affective video indexing and will guide the development of new corpora to avoid the shortcomings of existing ones. In the next sections, we first introduce the proposed specifications for affective video corpora and then we discuss how corpora that we have developed have move progressively towards addressing these limitations.

4 Specifications for Affective Video Corpora

In this section, we present a set of corpus development specifications for affective video indexing. The specifications are informed by the ground that we have covered thus far, i.e., understanding of emotions and affective response from psychology and techniques available to record it, and also by the general types of multimedia context analysis algorithms that we expect that researchers will be developing with the data sets. We also take into account the limitations of the currently existing corpora, just discussed. From this information, three dimensions emerge that are critical to take into consideration when developing corpora for affective video indexing.

Context of viewer emotional response: Emotional response is complex, and arises not just from the video, but from the context of the video. We consider context to be what the viewer was exposed to before and after the part of the video for which we are interested in the affective impact. Context also includes the people with whom the viewer is watching the video and the viewer's underlying mood and physical state. The complexity cannot be completely controlled, but its impact can be minimized by very explicitly planning the set up in which viewers are exposed to videos. An evaluation protocol should be included that describes exactly what the annotators were asked to do. The protocol ensures that the annotation situation is reproducible should it ever be necessary/desirable to extend the annotations. What is important is to remain firmly focused on how the task is defined so that it is clearly understood that we are trying to predict affective impact on the viewer. Modeling of affect expressed within the video (i.e., intrinsic affect) is admitted. But it should be understood that this is only used as a bridge to infer the ultimate impact on the viewer. It should be clearly stated, which parts of the emotional response process, for example, the affective and cognitive components vs. the conative and physiological components and it should be taken into consideration the implications of ignoring the other components.

The formulation of the way in which self-reported emotions are elicited should control for the impact of video before and after the target segment. This includes showing enough of the video. It is important to realize that entertainment video "works" exactly because it takes us as viewers through alternations of mood, or expresses more than one mood at once. Depending on how the video is split up for mood elicitations different (or impartial) viewer responses can be expected. Good handling of context also involves gathering information on the users' underlying mood and physical state.

Personal variation among viewers: Personal variation among viewers has a variety of sources. Some of the personal variation can be dealt with by careful handling of context, as mention above. Classical demographic differences are another source of variation. It is important that the target group be defined clearly (e.g., children) so that any existing limitations on user-to-user variability can be as well understood as possible. Narrowing the target group to a very small demographic (e.g., university students in their twenties) should be understood to limit the general applicability of the annotations gathered.

Personal reactions vary according to personal topic preference. It is important to abstract away from topic or the topical interest of viewers: this can be accomplished by using a well balanced data set. Alternately, during data set design, a decision can be made to focus on one particular topic or style of data, which is significant enough to merit study. In any case, the corpus should be as multifunctional as possible: for example, have enough users so that not only can universal reactions be studied, but also it is possible to study the reactions of different clusters of users that have similar responses.

Effectiveness and efficiency In practice, evaluation corpora are always developed under limitations of resources including person power and time. It is important to carefully plan how the corpus development process is handled. Decisions how to most effectively allocate limited resources have a critical effect on the usefulness of the corpus. As much as possible, such decisions should not be made in an arbitrary manner or during the actual process of gathering annotations for the corpus. In order to avoid unnecessarily jeopardizing the usefulness of the corpus, design decisions should be informed by the overall scenario or scenario for which the data set is being developed. An overall scenario is particularly helpful, should it become necessary to make further design decisions during the course of corpus development, e.g., decisions how to most effectively use limited resources in the case of unexpected loss of time or budget. Also, if researchers want to reuse the data set later, they have an idea of which uses are appropriate and which uses overstretch what the data set is designed to do. It is helpful to take into account the kind of multimedia content analysis that will be developed and the evaluation measure that will be used. However, it is of critical importance that the corpus be designed to reflect human affective reactions and not be biased to the specific algorithms, or types of algorithms, whose development it is intended to support.

5 Developed Corpora

To demonstrate the application of these specifications we now turn to discuss concrete examples. Three affective video corpora have been developed using three approaches of increasing sophistication, which progressively approach the ideal benchmarking corpus. The lessons learned from each corpus development experience were used to improve the next development. The annotations of the first dataset were gathered a laboratory setting. The second dataset was annotated with user affective responses gathered via a Web-based online platform and the third dataset includes affective responses gathered using an online crowdsourcing platform.

5.1 Movie scenes annotated in a laboratory environment

5.1.1 *Emotional videos*

The first corpus that was developed is comprised of emotional movie scenes suitable for emotion elicitation and characterization. The affective annotations were gathered via an experiment in which short video clips were shown to participants in a laboratory setting and their physiological responses and emotional self-reports were recorded. The results and analysis of physiological responses have been given in detail in [53]. Due to the limited time a participant can spend in each session, a relatively small set of videos, 64 clips from eight movies, were chosen and shown in two sessions. To create this video dataset, we selected video scenes from movies chosen either by following similar studies (e.g., [44], [54], [43]), or from recent popular movies. The set of movies included four major genres: drama, horror, action, and comedy, and are shown in in Table 2 together with the index codes they were assigned for use in the experiment. The scenes that were selected, eight for each movie, had durations of approximately one to two minutes each and contained an emotional event (as judged by the first author). The complete list of the scenes with editing instructions and descriptions is available online[1].

TABLE 2
Movies contained in the first corpus, organized by genre. Index number assigned to the movie is shown in parentheses after each title.

Drama movies	Comedy movies
The pianist (6), Hotel Rwanda (2)	Mr. Bean's holiday (5), Love actually (4)
Horror movies	**Action movies**
The ring (Japanese version) (7), 28 days later (1)	Kill Bill Vol. I (3), Saving private Ryan (8)

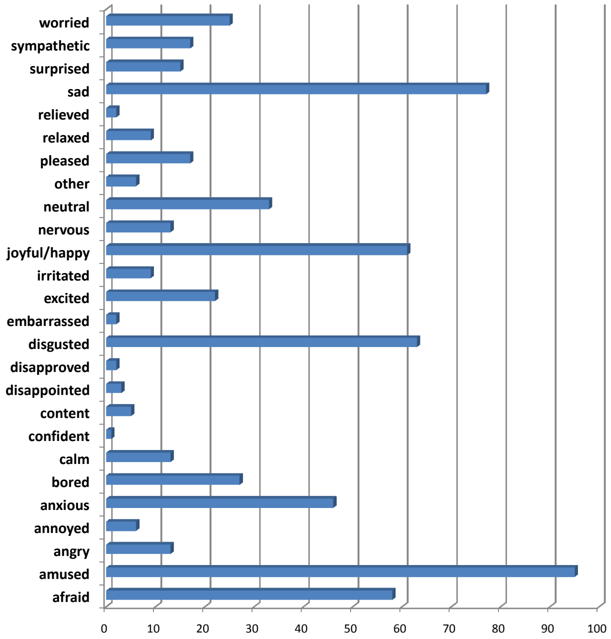


Fig. 4. Total number of keywords reported by 10 participants in response to the 64 video clips of movie scenes in the first corpus.

5.1.2 *Analysis of assessments*

The annotations were collected from ten (three female and seven male) participants ranging in age from 20 to 40 years ($M = 29.3, SD = 5.4$). The difference between arousal and valence scores given by the participants to all the videos was studied by means of a multi-way ANalysis Of VAriance (ANOVA), which was performed on arousal and valence scores considering three factors: the video scenes, the participants and the order in which the videos were shown to the participants during sessions. The effect of the order in which the videos were presented to the users on the user response was not significant. However, there was a significant

[1] http://cvml.unige.ch/movieList

difference on average valence scores between different participants ($F(9) = 18.53, p < 1 \times 10^{-5}$) and different videos ($F(63) = 12.17, p < 1 \times 10^{-5}$). There was also a significant difference on average arousal scores between different participants ($F(9) = 19.44, p < 1 \times 10^{-5}$) and different videos ($F(63) = 3.23, p < 1 \times 10^{-5}$). These differences can be attributed to different personal experiences and memories concerning different movies, as well as participants' mood and background.

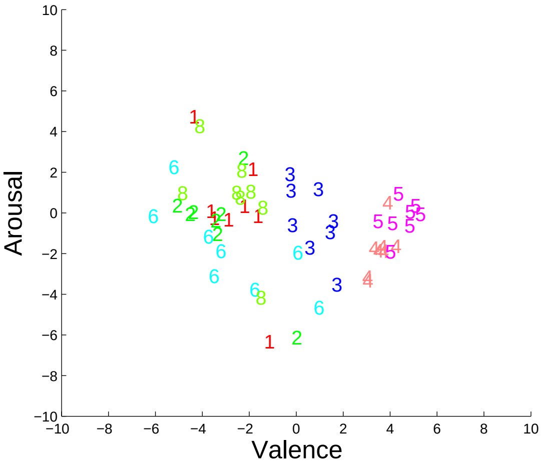
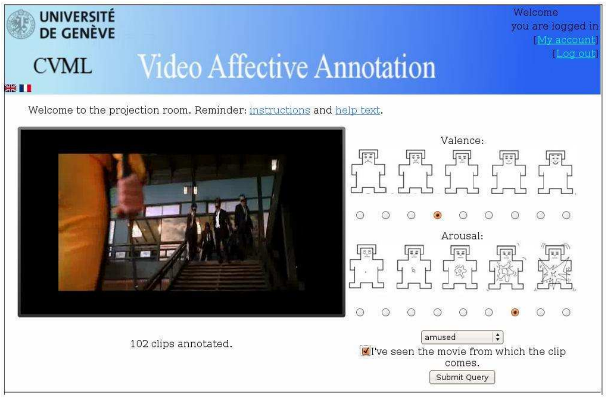


Fig. 5. The distribution of different movie scenes on arousal and valence plane. average arousal and valence are shown. Different numbers represent different movies (see Table 2).

The distribution of average arousal and valence scores are shown in Fig. 5. The numbers that represent the movie scenes are the codes associated with the movies in Table 2. The variance along the valence dimension was observed to increase with arousal. This observation is consistent with the findings of [57] in which arousal and valence scores in response to International Affective Picture System (IAPS) and International Affective Digital Sounds (IADS) showed a parabolic or heart shape distribution.

The development of this corpus provided an important lesson about the personal nature of user-reported affective response and the importance of carefully designing the method for collecting self-reported affective keywords from the participants. During the experiments the participants remarked that it was difficult for them to come up with emotion words when watching a video scene. In the end, there was a very low level of consensus among the words that they chose. The overall set of keywords chosen by the participants did not include a high number of instances of basic emotions, e.g., anger, was not very common compared to non-basic emotions, e.g., amusement (see Fig. 4). These observations led to the lesson that giving users complete freedom of choice of response will not help to isolate those common aspects of affective response. Instead, it is easier and yields more stable results if participants choose from a list of choices. However, the list should not be blindly adopted from the literature, but should be carefully developed for a particular setting using exploratory experiments with participants. Overall, the more participants that contribute affective response reports, the higher the consensus will be.



Video Affective Annotation

Welcome
you are logged in
[My account]
[Log out]

Welcome to the projection room. Reminder: instructions and help text.

Valence:

○ ○ ○ ● ○ ○ ○ ○ ○

Arousal:

○ ○ ○ ○ ○ ○ ● ○ ○

102 clips annotated.

amused

☑ I've seen the movie from which the clip comes.

Submit Query

Fig. 6. A snapshot of the affective annotation platform.

5.2 Web-based annotated movie scenes dataset

5.2.1 *Emotional videos*

The development of the second corpus targeted the involvement a larger set of participants. First, a user study was conducted to narrow down the selection of videos to be used as stimuli. This time a more efficient forced-choice self-reporting was used. In order to find videos eliciting emotions from the whole spectrum of possible emotions, a user study was conducted to annotate a set of manually preselected movie scenes. The dataset is drawn from 16 full length Hollywood movies (see [41] for the full list). To create this video dataset, we extracted video scenes from movies selected either according to similar studies (e.g., [44], [54], [43], [53]), or from recent famous movies. A set of 155 short clips, each about one to two minutes long, were manually selected from these movies to form the dataset.

A Web-based annotation system was developed and deployed in order to collected participants' self-reported affective responses. To use this system, experiment participants sign up and provide personal information including gender, age, and email address. The system also collects information such as cultural background and origin that is used to build a profile of the experiment participants. Providing this information is optional. Fig. 6 shows a screenshot of the assessment interface where a video clip is being shown. After watching each video clip, the participant expressed his/her felt emotion using arousal and valence, quantized in nine levels. Participants also choose the emotional label best reflecting the emotions that they feel upon watching the clips. The emotion labels are "afraid", "amused", "anxious", "disgusted", "joyful", "neutral", and "sad".

These labels have been chosen based on an assessment of the labels used in our first set of experiments (see Section 5.1).

5.2.2 *Analysis of the self-reports*

Initially, 82 participants signed up to annotate the videos. From these 82 participants, 42 participants annotated at least 10 clips. Participants were from 20 to 50 years old ($M = 26.9,\ SD = 6.1$). Out of the 42 participants, 27 were male and 15 were female with different cultural backgrounds living in four different continents. The results of a multi-way ANOVA on arousal scores as the dependent variable and participant, video clip, and time of day as effects showed that the average arousal scores have a significant difference for different participants ($F(41) = 3.23, p < 1 \times 10^{-5}$), video clips ($F(154) = 5.35, p < 1 \times 10^{-5}$) and time of day ($F(7) = 2.69, p < 0.01$). A day was divided into eight time interval, early morning (6:00 to 9:00), morning (9:00 to 11:30), noon (11:30 to 13:00), afternoon (13:00 to 16:30), evening (16:30 to 19:30), late evening (19:30 to 22:30), night (22:30 to 24:00) and after midnight (00:00 to 6:00). The average arousal scores in different time periods are shown in Fig. 7. The average arousal scores given to all videos increases from early in the morning until noon. Then it decreases until it bounces back for late evening and night. The effect of circadian rhythm on self-reported arousal levels reflects the impact of context. Female participants on average gave higher arousal scores to the videos. A Wilcoxon rank sum test showed that the difference between female and male participants' arousal scores was significant ($p = 3 \times 10^{-16}$). These results are in line with the previous findings, e.g., [54], which showed women report stronger emotions than men in response to the same stimuli.

Using the Web-based annotation system had a clear advantage because it increased the number of users from which we were able to collect annotations. We were able to collect enough annotations so that it was meaningful to analyze the variance of the reported arousal scores. Also, the Web interface meant that the participants could annotate more video than what was possible in two lab sessions of limited length. Additionally, the development of this corpus led to an important lesson about the context of user annotations. Using the Web interface meant that the environment of the affective response was less controlled. The time of day had a significant impact on the response and it was important to record information about the influence of this factor.

5.3 Boredom prediction dataset

5.3.1 *Crowdsourcing for affective annotation*

In order to reach a broader, more diverse, and larger population, a crowdsourcing platform, Amazon Mechanical Turk (MTurk)[2], was used to gather annotations in the development third dataset. The third dataset, initially described in [58], was developed with the aim of supporting research on video processing algorithms capable of predicting viewer boredom. A video dataset has been gathered in the context of the MediaEval[3] 2010 Affect Task for boredom prediction of Internet videos. Using MTurk we rapidly gathered self-reported boredom scores from a large user group that is demographically diverse and also represented our target population (Internet video viewers). Again, the forced choice emotional self-reporting methods were employed.

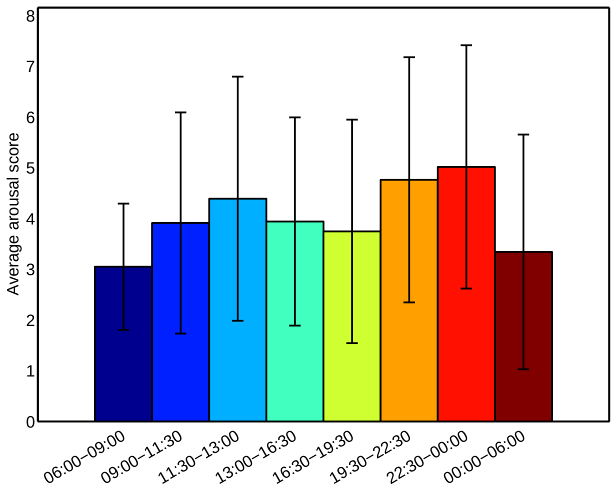


Fig. 7. Average arousal scores in different times of the day.

For this work, we adopted a relatively simple, straightforward definition of viewer-experienced boredom. Boredom was taken to be related to the viewer's sense of maintaining focus of attention and is related to the apparent passage of time [59]. Boredom is understood to be a negative feeling associated with viewer perceptions of the viewer-perceived quality (i.e., viewer appeal) of the video being low.

The dataset selected for the corpus is Bill's Travel Project, a travelogue series called "My Name is Bill" created by the film maker Bill Bowles[4]. The series consists of 126 videos between two to five minutes in length. This data set was chosen since it represents the sort of multimedia content that has risen to prominence on the Web. Bill's travelogue follows the format of a daily episode related to his activities and as such is comparable to "video journals" that are created by many video bloggers.

5.3.2 *Design of crowsdsourcing task*

The third corpus that was developed once again increased the number of annotators and also introduced an even more sophisticated mechanism for context control. The affective responses for this corpus where collected using a large commercial crowdsourcing platform,

[2] http://www.mturk.com

[3] http://www.multimediaeval.org

[4] http://www.mynameisbill.com

Amazon Mechanical Turk (http://www.mturk.com). A crowdsourcing platform is an online labor market in which microtasks are offered by requestors and carried out by a pool of human users referred to as “workers”. Work in the area of human judgment and decision-making has revealed that there is no difference in the magnitude of the observed effects when experiments are performed using Mechanical Turk and when they are performed with a conventional pool of subjects [60].

The crowdsourcing strategy used for the third corpus was designed based on the existing crowdsourcing literature, for example [61], online articles and blog posts about crowdsourcing such as “Behind the enemy lines” blog[5], and also taking into account our past experience regarding collecting annotations in the Web-based experiment. A two-step approach was taken for our data collection. The first step was the pilot that consisted of a single micro-task or Human Intelligent Task (HIT) involving one video. This first HIT was used for the purpose of recruiting and screening MTurk workers as experiment participants. The second step was the main task and involved a series of 125 micro-tasks, one for each of the remaining videos in the collection. Workers were paid 30 US dollar cents for each HIT that they successfully completed.

The pilot HIT contained three components corresponding to responses that were required from the experiment participants that we recruited. The first section contained questions about the personal background (i.e., age, gender, cultural background). The second section contained questions about viewing habits: workers were asked whether they were regular viewers of Internet videos. The third section tested their seriousness by asking them to watch the video, select a word that reflected their mood at the moment and also write a summary. The summary constituted a “verifiable” question, recommended by [61]. The summary offered several possibilities for verification. Its length and whether it contained well-formulated sentences gave us an indication of the level of care that the worker devoted to the HIT. Also, the descriptive content indicated whether the worker had watched the entire video, or merely the beginning. A final question inquired if they were interested in performing further HITs of the same sort. In order to hide the main goal of the study from workers, the text box for the video summary was placed prominently in the HIT.

The workers were chosen and qualifications were granted for the main task from the participants of the pilot by considering the quality of their description and answers. In the choice of workers, we also strove to maintain a diverse group of respondents. Each HIT in the main study consisted of three parts. In the first part, the workers were asked to specify the time of day. Also the workers were asked to choose a mood word from a drop down list that best expressed their reaction to an imaginary word (i.e., a nonsense word), such as those used in [62]. The mood words were “pleased”, “helpless”, “energetic”, “nervous”, “passive”, “relaxed”, and “aggressive”. The answers to these questions gave us an estimate of their underlying mood. In the second part, they were asked to watch the video and give some simple responses to the following questions. They were asked to choose the word that best represented the emotion they felt while watching a video from a second list of emotion words in the drop down list. The emotion list contained the Ekman six basic emotions [55] (namely, “sadness”, “joy”, “anger”, “fear”, “surprise”, and “disgust”) in addition to “boredom”, “anxiety”, “neutral” and “amusement”, which cover the entire affective space, as defined by the conventional dimensions of valence and arousal [29]. The emotion and mood word lists contained different items in order to avoid as much as possible that the experiment participants would strongly associate the two. Next, they were asked to provide a rating specifying how boring they found the video and how much they liked the video, both on a nine point scale. Finally, they were asked to describe the contents of the video in one sentence.

5.3.3 Analysis of the ratings

Our pilot HIT was initially published for 100 workers and finished in the course of a single weekend. We republished the HIT for more workers when we realized we needed more people in order to have an adequate number of task participants. Only workers with the HIT acceptance rate of 95% or higher were admitted to participate in the pilot HIT. In total, 169 workers completed our pilot HIT, 87.6% of which reported that they watch videos on the Internet. We took this response as confirmation that our tasks participants were close to the target audience of our research. Out of 169 workers, 105 were male and 62 were female and two did not report their gender. Their age average was 30.48 with the standard deviation of 12.39. The workers in the pilot HITs identified themselves by different cultural backgrounds from North America. Having such a group of participants with a high diversity in their cultural background would have been difficult in a conventional setting, i.e., without using the crowdsourcing platform. Of the 169 pilot participants, 162 had interest in carrying out similar HITs. Out of the interested group, 79 workers were determined to be qualified and were assigned our task-specific qualification within MTurk. This means only 46.7% of the workers who did the pilot HIT were able to answer all the questions and had the profile we required for the main task.

In total, 32 workers participated and also annotated more than 60 of the 125 videos in the main task HIT series. This means only 18.9% of the participants in the pilot and 39.0% of the qualified participants committed to do the main task HIT series seriously. Of this group of 32 serious participants, 18 were male and 11 were female with ages ranging from 18 to 81 ($M = 34.9, SD = 14.7$).

[5] http://behind-the-enemy-lines.blogspot.com

The following questions were asked about each video to assess the level of boredom. First, how boring the video was on nine-point scale from the most to the least boring. Second, how much the user liked the video on the nine-point scale and third how long the video was. Boredom was shown to have on average a strong negative correlation, $\rho = -0.86$ with liking scores. The correlation between the order of watching the videos for each participant and the boredom ratings was also examined. No positive linear correlation was found between the order and boredom score. This means that watching more videos did not increase the level of boredom and, in fact, for two of the participants it lowered their reported boredom level. Additionally, the correlation between the video length and boredom scores was investigated. No positive correlation was found between the boredom scores and videos' duration. We can conclude that longer videos are not necessarily perceived as more boring than the shorter videos.

To measure the inter-annotator agreement, the Spearman correlation between participants' pairwise boredom scores was computed. The average significant correlation coefficient was very low $\rho = 0.05$. There were even cases where the correlation coefficients were negative, which shows complete disagreement between participants. The low inter-annotators agreement reflects the personal taste in boredom perception. However, The rank for average boredom scores were robust among the extreme cases and reproducible with a subset of users.

For each worker we then grouped videos into two rough categories, above and below the mean boredom score of that worker. We computed the average pair-wise Cohen's kappa for these categories and here found a slight agreement ($\kappa = 0.01 \pm 0.04$). We also compared agreement on the emotion words workers associated with viewers. Here, again Cohen's kappa indicated a slight agreement ($\kappa = 0.05 \pm 0.06$). The weak correlations suggest that it is indeed important to investigate personalized approaches to affective response prediction.

In order to obtain the dominant mood from the mood words, first the responses of each participant were clustered into the three hours time intervals. In each three hours interval the most frequent chosen mood word was selected as the dominant mood. After calculating the dominant moods, we found that using the implicit mood assessment none of the participants had the "relaxed" as their dominant mood.

The average boredom scores for different dominant moods are shown in Fig. 8. The boredom scores were, on average, lower, i.e., indicated that videos were *more* boring, for viewers in a passive mood and higher, i.e., indicated that videos were *less* boring, in an arguably more active active mood such as "energetic", "nervous" and "pleased". Moods were then categorized into positive ("pleased", "energetic", and "relaxed") and negative, ("helpless", "nervous", "passive", and "aggressive") categories. On average, participants gave higher ratings to videos when they were in positive moods.The statistical significance of the difference between ratings in positive and negative moods was examined by a Wilcoxon test and was found significant ($p = 4 \times 10^{-8}$). The effect of four different factors on boredom scores was investigated with a four-way ANOVA. The effects were mood, time of day, videos and participants. Unlike the second experiment, the effect of the time of day on boredom scores was not significant. This observation can be attributed to the difference between the nature of arousal and perceived boredom, arousal being more correlated with physiological state. Participants' mood had a significant effect on the ratings ($F(6) = 5.55, p < 1 \times 10^{-4}$). The interaction between each pair of actors was investigated to check whether the observed difference was as a result of having special videos for every mood. The interaction in two-way ANOVA between the videos and moods was not significant. Therefore, the effect of mood on boredom scores was independent from the effect of videos. The analysis of annotations gathered in this dataset showed the importance of participants' mood which is often not assessed in affective and non-affective assessments.

From the development of this corpus we learned that it is possible to use commercial crowdsourcing to collect a large volume of user affective responses to video. The large number of participants made it possible to analyze sub-groups of participants with particular reactions. Affective response is personal, and varies from individual. However, looking at sub-groups of the population that pattern in the same way could make it possible to isolate the commonalities of viewer response. Variations in the context were addressed by collecting information on the underlying moods of the participants. Even though crowdsourcing is relatively inexpensive, it is still important to plan resources carefully when designing a corpus that uses crowdsourcing to collect affective annotations. A trade-off needs to be made between more annotators, the number of videos annotated and the parts of the HIT design that verify engagement. Advanced planning was necessary to collect the annotations within a set amount of time, since workers may not immediately start working on a HIT once they have qualified. Also, it was necessary to have a large enough pool of workers available to work on the HIT, since some qualified workers do not return to work on the HIT after earning the qualification. In sum, the third corpus addressed all three dimensions of personal affective response, context and tradeoffs of efficiency and effectiveness.

6 Conclusion and Outlook

This paper has addressed the development of corpora for research and evaluation in the area of affective video indexing algorithms and systems. We have proposed a set of specifications that are intended to provide the research community with support in overcoming the deficiencies of the existing corpora for affective video indexing. Our investigation focused not only on the

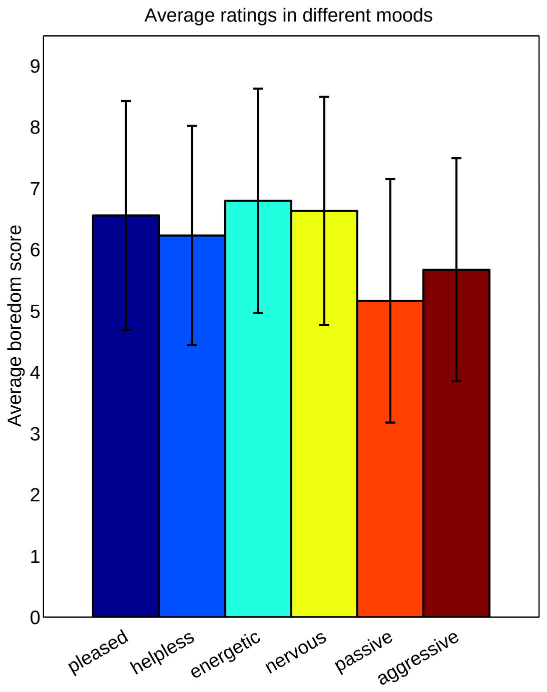


Fig. 8. Average dominant boredom scores reported by viewers experimencing different moods.

sources of the size- and scope-related limitations of the existing evaluation corpora, namely the difficulty of reliably collecting affective responses from test users and the need to reduce the variability in the noisy and subjective affective responses, but also on how other critical requirements related to the corpora development can be fulfilled.

Our findings indicate that there are three key dimensions that need to be considered when developing corpora for affective video indexing research. The first dimension is context. In particular, circadian rhythm and the mood of experiment participants have been shown to have significant effect on self-reported emotions. The second dimension is the personal differences. Significant differences have been observed between different participants' affective responses to the same content. The personal dimension emphasizes the importance of personalization and profiling strategies. Finally, efficiency and effectiveness are important factors to be taken into account. New methods for collecting annotations such as Web-based and crowdsourcing platforms offer improved opportunities to collect annotations in greater volumes and from wider diversity of users and a broader spectrum of contexts.

These three dimensions form the basis for our proposed set of specifications for affective indexing corpora. Three corpora are introduced which are developed with techniques that progressively approach an ideal corpus as defined by these specifications. Forced-choice and simple emotional reporting methods have been employed in developing the corpora to decrease the effort of the participants and increase the efficiency of annotation collection and the ability of the annotations to reflect consistency and stability in self-reported affective responses. Larger populations of experimental participants can be reached with Web-based and crowdsourcing platforms, increasing both the diversity of the annotations collected and the ability of the corpus to reflect contextual and personal variation.

This paper has argued that high-quality corpora will help to push forward the state of the art in affective video indexing. In order to realize this predicted potential of affective video corpora, the next step is necessarily the development of additional corpora. If a multitude of corpora can be made available to the research community suitable to support research along the entire spectrum of possible affective video indexing applications, then researchers will have the necessary resources at their disposal to push affective video indexing into the next generation.

Acknowledgments

The work of Larson and Hanjalic is supported in part by the European Community's Seventh Framework Program under grant agreement no. 287704 (CUbRIK). This work of Soleymani is supported by the European Research Council under the FP7 Marie Curie Intra-European Fellowship: Emotional continuous tagging using spontaneous behavior (EmoTag). The authors would like to thank Bill Bowles for granting permission of using his video content for our boredom detection corpus.

References

[1] H.-B. Kang, "Affective content detection using HMMs," in *Proceedings of the eleventh ACM international conference on Multimedia*, ser. MULTIMEDIA '03. New York, NY, USA: ACM, 2003, pp. 259–262.

[2] A. Hanjalic, "Adaptive Extraction of Highlights From a Sport Video Based on Excitement Modeling," *Multimedia, IEEE Transactions on*, vol. 7, no. 6, pp. 1114–1122, 2005.

[3] C. H. Chan and G. J. F. Jones, "Affect-based indexing and retrieval of films," in *MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia*. New York, NY, USA: ACM, 2005, pp. 427–430.

[4] C. G. M. Snoek and M. Worring, "Concept-based video retrieval," *Foundations and Trends in Information Retrieval*, vol. 4, no. 2, pp. 215–322, 2009.

[5] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and trecvid," in *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*. New York, NY, USA: ACM Press, 2006, pp. 321–330.

[6] ——, "High-Level Feature Detection from Video in TRECVid: a 5-Year Retrospective of Achievements," in *Multimedia Content Analysis, Theory and Applications*, A. Divakaran, Ed. Berlin: Springer Verlag, 2009, pp. 151–174.

[7] I. Arapakis, J. M. Jose, and P. D. Gray, "Affective feedback: an investigation into the role of emotions in the information seeking process," in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR '08. New York, NY, USA: ACM, 2008, pp. 395–402.

[8] I. Lopatovska and I. Arapakis, "Theories, methods and current research on emotions in library and information science, information retrieval and human-computer interaction," *Information Processing & Management*, vol. 47, no. 4, pp. 575 – 592, 2011.

[9] R. W. Picard, “Affective computing,” MIT, Media Laboratory Perceptual Computing Section Technical Report 321, November 1995.

[10] A. Hanjalic, “Extracting moods from pictures and sounds: towards truly personalized tv,” *Signal Processing Magazine, IEEE*, vol. 23, no. 2, pp. 90 –100, march 2006.

[11] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain, “Content-based multimedia information retrieval: State of the art and challenges,” *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 2, pp. 1–19, February 2006.

[12] S. Benini, L. Canini, and R. Leonardi, “A connotative space for supporting movie affective recommendation,” *Multimedia, IEEE Transactions on*, vol. PP, no. 99, p. 1, 2011.

[13] C. V. Thornley, A. C. Johnson, A. F. Smeaton, and H. Lee, “The scholarly impact of trecvid (2003–2009),” *J. Am. Soc. Inf. Sci. Technol.*, vol. 62, pp. 613–627, April 2011.

[14] Y. Rui, A. Gupta, and A. Acero, “Automatically extracting highlights for TV baseball programs,” in *Proceedings of the eighth ACM international conference on Multimedia*, ser. MULTIMEDIA ’00. New York, NY, USA: ACM, 2000, pp. 105–115.

[15] G. J. Jones and C. Hau-Chan, *Affect-Based Indexing for Multimedia Data*. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2012.

[16] A. Janin, L. Gottlieb, and G. Friedland, “Joke-o-mat HD: browsing sitcoms with human derived transcripts,” in *Proceedings of the international conference on Multimedia*, ser. MM ’10. New York, NY, USA: ACM, 2010, pp. 1591–1594.

[17] W. Wirth and H. Schramm, “Media and Emotions,” *Communication research trends*, vol. 24, no. 3, pp. 3–39, 2005.

[18] K. R. Scherer, “What are emotions? And how can they be measured?” *Social Science Information*, vol. 44, no. 4, pp. 695–729, December 2005.

[19] ——, “Studying the emotion-antecedent appraisal process: An expert system approach,” *Cognition & Emotion*, vol. 7, no. 3, pp. 325–355, 1993.

[20] A. Ortony, G. L. Clore, and A. Collins, *The Cognitive Structure of Emotions*. Cambridge University Press, July 1988.

[21] D. Sander, D. Grandjean, and K. R. Scherer, “A systems approach to appraisal mechanisms in emotion,” *Neural Netw.*, vol. 18, no. 4, pp. 317–352, May 2005.

[22] D. Zillmann, *The psychology of suspense in dramatic exposition*. Lawrence Erlbaum Associates, Inc, 1996, pp. 199–231.

[23] A. I. Nathanson, *Rethinking Empathy*. Lawrence Erlbaum Associates, Inc, 2003, ch. 5, pp. 107–130.

[24] D. Zillmann, *Empathy: Affect from bearing witness to the emotions of others*. Lawrence Erlbaum Associates, Inc, 1991, pp. 135–168.

[25] P. Ekman, *Basic Emotions*. John Wiley & Sons, Ltd, 2005, pp. 45–60.

[26] W. Wundt, *Grundzüge der physiologischen Psychologie*. Leipzig: Engelmann, 1905.

[27] S. Marsella, J. Gratch, and P. Petta, *Computational models of emotion*. Oxford, UK: Oxford University Press, 2010, ch. 1.2, pp. 21–41.

[28] J. A. Russell, “Culture and the Categorization of Emotions,” *Psychological Bulletin*, vol. 110, no. 3, pp. 426–450, 1991.

[29] J. A. Russell and A. Mehrabian, “Evidence for a three-factor theory of emotions,” *Journal of Research in Personality*, vol. 11, no. 3, pp. 273–294, September 1977.

[30] J. R. J. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth, “The World of Emotions is not Two-Dimensional,” *Psychological Science*, vol. 18, no. 12, pp. 1050–1057, 2007.

[31] P. Desmet, *Measuring emotion: development and application of an instrument to measure emotional responses to products*. Norwell, MA, USA: Kluwer Academic Publishers, 2003, ch. 9, pp. 111–123.

[32] P. Winoto and T. Y. Tang, “The role of user mood in movie recommendations,” *Expert Systems with Applications*, vol. 37, no. 8, pp. 6086–6092, 2010.

[33] M. M. Bradley and P. J. Lang, “Measuring emotion: the Self-Assessment Manikin and the Semantic Differential.” *J Behav Ther Exp Psychiatry*, vol. 25, no. 1, pp. 49–59, March 1994.

[34] J. A. Russell, A. Weiss, and G. A. Mendelsohn, “Affect Grid: A single-item scale of pleasure and arousal,” *Journal of Personality and Social Psychology*, vol. 57, no. 3, pp. 493–502, September 1989.

[35] J. A. Russell, “A circumplex model of affect.” *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, December 1980.

[36] A. Schaefer, F. Nils, X. Sanchez, and P. Philippot, “Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers,” *Cognition & Emotion*, vol. 24, no. 7, pp. 1153–1172, 2010.

[37] D. Watson and L. A. Clark, “The PANAS-X: Manual for the Positive and Negative Affect Schedule–Expanded Form,” 1994.

[38] O. Villon, “Modeling affective evaluation of multimedia contents: user models to Associate subjective experience, physiological expression and contents description,” Ph.D. dissertation, Université de Nice - Sophia Antipolis, Nice, France, October 2007.

[39] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. Mcmahon, M. Sawey, and M. Schröder. (2000) feeltrace’: an instrument for recording perceived emotion in real time.

[40] E. Douglas-Cowie, R. Cowie, and M. Schröder, “A New Emotion Database: Considerations, Sources and Scope,” in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000, pp. 39–44.

[41] M. Soleymani, J. Davis, and T. Pun, “A collaborative personalized affective video retrieval system,” in *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, sep 2009.

[42] C.-H. Demarty, C. Penet, G. Gravier, and M. Soleymani, “A benchmarking campaign for the multimodal detection of violent scenes in movies,” in *ECCV Workshops (3)*, ser. Lecture Notes in Computer Science, A. Fusiello *et al.*, Eds., vol. 7585. Springer, 2012, pp. 416–425.

[43] A. Hanjalic and L.-Q. Xu, “Affective video content representation and modeling,” *Multimedia, IEEE Transactions on*, vol. 7, no. 1, pp. 143–154, 2005.

[44] H. L. Wang and L.-F. Cheong, “Affective understanding in film,” *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 16, no. 6, pp. 689–704, jun 2006.

[45] S. Arifin and P. Cheung, “Affective Level Video Segmentation by Utilizing the Pleasure-Arousal-Dominance Information,” *Multimedia, IEEE Transactions on*, vol. 10, no. 7, pp. 1325–1341, 2008.

[46] M. Xu, J. S. Jin, S. Luo, and L. Duan, “Hierarchical movie affective content analysis based on arousal and valence features,” in *Proceeding of the 16th ACM international conference on Multimedia*, ser. MM ’08. New York, NY, USA: ACM, 2008, pp. 677–680.

[47] M. Soleymani, J. J. M. Kierkels, G. Chanel, and T. Pun, “A Bayesian framework for video affective representation,” in *Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII 2009)*, September 2009, pp. 1–7.

[48] G. Irie, T. Satou, A. Kojima, T. Yamasaki, and K. Aizawa, “Affective Audio-Visual Words and Latent Topic Driving Model for Realizing Movie Affective Scene Classification,” *Multimedia, IEEE Transactions on*, vol. 12, no. 6, pp. 523–535, October 2010.

[49] H. Joho, J. Staiano, N. Sebe, and J. Jose, “Looking at the viewer: analysing facial activity to detect personal highlights of multimedia contents,” *Multimedia Tools and Applications*, vol. 51, no. 2, pp. 505–523, October 2010.

[50] R. M. Teixeira, T. Yamasaki, and K. Aizawa, “Determination of emotional content of video clips by low-level audiovisual features,” *Multimedia Tools and Applications*, pp. 1–29, Jan. 2011.

[51] H. Joho, J. M. Jose, R. Valenti, and N. Sebe, “Exploiting facial expressions for affective video summarisation,” in *Proceeding of the ACM International Conference on Image and Video Retrieval*, ser. CIVR ’09. New York, NY, USA: ACM, 2009.

[52] J. J. M. Kierkels, M. Soleymani, and T. Pun, “Queries and tags in affect-based multimedia retrieval,” in *ICME’09: Proceedings of the 2009 IEEE international conference on Multimedia and Expo*. Piscataway, NJ, USA: IEEE Press, 2009, pp. 1436–1439.

[53] M. Soleymani, G. Chanel, J. J. M. Kierkels, and T. Pun, “Affective Characterization of Movie Scenes Based on Content Analysis and Physiological Changes,” *International Journal of Semantic Computing*, vol. 3, no. 2, pp. 235–254, June 2009.

[54] J. Rottenberg, R. D. Ray, and J. J. Gross, *Emotion elicitation using films*, ser. Series in affective science. Oxford University Press, 2007, pp. 9–28.

[55] P. Ekman, W. V. Friesen, M. O’Sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W. A. LeCompte, T. Pitcairn, and P. E. Ricci-Bitti, “Universals and cultural differences in the judgments of facial expressions of emotion.” *Journal of personality and social psychology*, vol. 53, no. 4, pp. 712–717, October 1987.

[56] R. Plutchik, *A general psychoevolutionary theory of emotion*. New York: Academic press, 1980, pp. 3–33.

[57] R. B. Dietz and A. Lang, “Aefective agents: Effects of agent affect on arousal, attention, liking and learning,” in *Cognitive Technology Conference*, 1999.

[58] M. Soleymani and M. Larson, “Crowdsourcing for affective annotation of video: Development of a viewer-reported boredom

corpus," in *Workshop on Crowdsourcing for Search Evaluation, SIGIR 2010*, Geneva, Switzerland, July 2010.

[59] J. D. Laird, *Feelings: The Perception of Self*, 1st ed. USA: Oxford University Press, Jan. 2007.

[60] G. Paolacci, J. Chandler, and P. G. Ipeirotis, "Running Experiments on Amazon Mechanical Turk," *Judgment and Decision Making*, vol. 5, no. 5, pp. 411–419, Aug. 2010.

[61] A. Kittur, E. H. Chi, and B. Suh, "Crowdsourcing user studies with Mechanical Turk," in *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, ser. CHI '08. New York, NY, USA: ACM, 2008, pp. 453–456.

[62] M. Quirin, M. Kazén, and J. Kuhl, "When Nonsense Sounds Happy or Helpless: The Implicit Positive and Negative Affect Test (IPANAT)," *Journal of Personality and Social Psychology*, vol. 97, no. 3, pp. 500–516, 2009.